My Analysis Toolbox

Tri Nguyen

nguyetr9@oregonstate.edu

August 18, 2023

In progress . . .

This is a collection of what I have used in my works, or seen from other interesting works. This should serve as a warehouse for me to casually browser through when trying to find ideas. In order to avoid cluttering, all the proofs are moved to the end of this note.

Contents

List of Theorems

1 Tensor

1.1 Khatri-Rao product

Lemma 1.1 (Khatri-Rao product)**.** *vec* $\left(\boldsymbol{A}\boldsymbol{D}\boldsymbol{B}^T\right) = (\boldsymbol{B} \odot \boldsymbol{A})\boldsymbol{d} \quad$ *where* $\boldsymbol{D} = Diag(\boldsymbol{d})$

1.2 Special case of lemma permutation

Lemma 1.2 (Special case of lemma permutation)**.** *Given 2 nonsingular matrice* $\overline{\mathbf{C}}, \mathbf{C} \in \mathbb{R}^{n\times n}$*. If* $w(\mathbf{v}^T\overline{\mathbf{C}}) = 1$ *implies* $w(\mathbf{v}^T\mathbf{C}) = 1$*, then*

$$\overline{\mathbf{C}} = \mathbf{C}\Pi\Lambda$$

Proof. We have

$$\bar{\mathbf{C}}^{-1}\bar{\mathbf{C}} = \begin{bmatrix} \mathbf{v}_1^T \\ \mathbf{v}_2^T \\ \dots \\ \mathbf{v}_n^T \end{bmatrix} \bar{\mathbf{C}} = \mathbf{I}$$

Since

$$w(\mathbf{v}_i^T\bar{\mathbf{C}}) = 1 \Rightarrow w(\mathbf{v}_i^T\mathbf{C}) = 1 \tag{1}$$

And because $\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_n$ are linearly independent and $\mathbf{C}$ is nonsingular,

$$\mathbf{v}_1^T\mathbf{C}, \mathbf{v}_2^T\mathbf{C}, ..., \mathbf{v}_n^T\mathbf{C} \text{ are linearly independent} \tag{2}$$

From 1, 2

$$\Rightarrow \bar{\mathbf{C}}^{-1}\mathbf{C} = \Pi^T\mathbf{D} \Rightarrow\Rightarrow \overline{\mathbf{C}} = \mathbf{C}\mathbf{D}^{-1}\Pi$$

■

1.3 CPD uniqueness - Simple case

Lemma 1.3 (CPD uniqueness - Simple case)**.** *Given* $\mathcal{X} = [[\mathbf{A}, \mathbf{B}, \mathbf{C}]]$ *where* $\mathcal{X} \in \mathbb{R}^{I\times J\times 2}, \mathbf{A} \in \mathbb{R}^{I\times F}, \mathbf{B} \in \mathbb{R}^{J\times F}, \mathbf{C} \in \mathbb{R}^{2\times F}$*. If* $k_{\mathbf{C}} = 2$ *and* $r_{\mathbf{A}} = r_{\mathbf{B}} = F$ *then the decomposition of* $\mathcal{X}$ *is essential unique.*

Proof. Two slabs of $\mathcal{X}$ are:

$$\mathcal{X}^{(1)} = \mathcal{X}(:,:,1) = \mathbf{A}\mathbf{D}_1(\mathbf{C})\mathbf{B}^T$$
$$\mathcal{X}^{(2)} = \mathcal{X}(:,:,2) = \mathbf{A}\mathbf{D}_2(\mathbf{C})\mathbf{B}^T$$

Define $\widetilde{\mathbf{A}} = \mathbf{A}\mathbf{D}_1(\mathbf{C}), \mathbf{D} = \mathbf{D}_1(\mathbf{C})^{-1}\mathbf{D}_2(\mathbf{C})$

$$\Rightarrow \mathcal{X}^{(1)} = \widetilde{\mathbf{A}}\mathbf{B}^T$$
$$\mathcal{X}^{(2)} = \widetilde{\mathbf{A}}\mathbf{D}\mathbf{B}^T$$
$$\Rightarrow \overline{\mathcal{X}} = \begin{bmatrix} \mathcal{X}^{(1)} \\ \mathcal{X}^{(2)} \end{bmatrix} = \begin{bmatrix} \widetilde{\mathbf{A}} \\ \widetilde{\mathbf{A}}\mathbf{D} \end{bmatrix} \mathbf{B}^T$$
$$\Rightarrow \mathcal{R}(\mathcal{X}) = \mathcal{R}\left(\begin{bmatrix} \widetilde{\mathbf{A}} \\ \widetilde{\mathbf{A}}\mathbf{D} \end{bmatrix}\right) \quad \text{since } \mathbf{B}^T \text{ is full row rank} \tag{3}$$

Meanwhile, apply SVD to $\begin{bmatrix} \mathcal{X}^{(1)} \\ \mathcal{X}^{(2)} \end{bmatrix}$, we obtain:

$$
\begin{aligned}
\text{thin svd}\left(\begin{bmatrix}\mathcal{X}^{(1)}\\ \mathcal{X}^{(2)}\end{bmatrix}\right) &= \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T \\
\Rightarrow \mathcal{R}(\overline{\mathcal{X}}) = \mathcal{R}(\mathbf{U}) &= \mathcal{R}\left(\begin{bmatrix}\mathbf{U}_1\\ \mathbf{U}_2\end{bmatrix}\right) \qquad (4)
\end{aligned}
$$

From 3 and 4, there exist a nonsingular matrix $\mathbf{M} \in \mathbb{R}^{F\times F}$:

$$
\begin{bmatrix}\mathbf{U}_1\\ \mathbf{U}_2\end{bmatrix} = \begin{bmatrix}\widetilde{\mathbf{A}}\\ \widetilde{\mathbf{A}}\mathbf{D}\end{bmatrix}\mathbf{M}
$$

Define

$$
\begin{aligned}
\mathbf{R}_1 &= \mathbf{U}_1^T\mathbf{U}_1 = \mathbf{M}^T\widetilde{\mathbf{A}}^T\widetilde{\mathbf{A}}\mathbf{M} = \mathbf{Q}\mathbf{M} \\
\mathbf{R}_2 &= \mathbf{U}_1^T\mathbf{U}_2 = \mathbf{M}^T\widetilde{\mathbf{A}}^T\widetilde{\mathbf{A}}\mathbf{D}\mathbf{M} = \mathbf{Q}\mathbf{D}\mathbf{M}
\end{aligned}
$$

They have similar form with $\mathbf{U}_1, \mathbf{U}_2$ except they are square, and nonsigular. Thus,

$$
\mathbf{R}_1\mathbf{R}_2^{-1} = \mathbf{Q}\mathbf{D}^{-1}\mathbf{Q}^{-1} \Rightarrow \mathbf{R}_2\mathbf{R}_1^{-1} = \mathbf{Q}\mathbf{D}\mathbf{Q}^{-1}
$$

$\mathbf{R}_2\mathbf{R}_1^{-1}$ is eigendecomposed to $\mathbf{Q}$ and $\mathbf{D}$. Therefore, we can find $\mathbf{D}, \mathbf{Q}$ by eigendecomposition of $\mathbf{R}_1\mathbf{R}_2^{-1}$. These 2 matrices are unique, but up to scale and permutation. What we have found are:

$$
\overline{\mathbf{Q}} = \mathbf{Q}\Pi\Lambda
$$

Back substitution to find $\widetilde{\mathbf{A}}$, then $\mathbf{A}, \mathbf{B}, \mathbf{C}$. All these matrices are unique but up to scale and permutation. That completes the proof. ∎

2 Matrix Algebra

2.1 Grammian matrix

Lemma 2.1 (Gramian matrix)**.** *If* $\mathbf{A} \in \mathbb{R}^{m\times n}$ *is full column rank, then* $\mathbf{A}^T\mathbf{A}$ *is invertible.*

Proof. Since $\mathbf{A}$ is full column rank, then $\mathbf{A}\mathbf{x} = \mathbf{0} \Leftrightarrow \mathbf{x} = \mathbf{0}$

$$
\begin{gathered}
\Rightarrow \mathbf{x}^T\mathbf{A}^T\mathbf{A}\mathbf{x} = \|\mathbf{A}\mathbf{x}\|_2^2 > 0 \quad \text{for all } \mathbf{x} \neq \mathbf{0} \\
\Rightarrow \mathbf{A}^T\mathbf{A}\mathbf{x} \neq \mathbf{0} \quad \text{for all } \mathbf{x} \neq \mathbf{0} \\
\Rightarrow \mathbf{A}^T\mathbf{A} \quad \text{is full rank} \Rightarrow \mathbf{A}^T\mathbf{A} \quad \text{is invertible}
\end{gathered}
$$

Comments:

- Contradictary proof will be easier

∎

2.2 Rank/Range of matrix multiplication

Lemma 2.2 (Rank/Range of matrix multiplication)**.** *Given* $\mathbf{A} \in \mathbb{R}^{m\times n}, \mathbf{B} \in \mathbb{R}^{n\times p}$*, and* $\mathbf{B}$ *is full row rank, then:*

$$
\mathcal{R}(\mathbf{A}\mathbf{B}) = \mathcal{R}(\mathbf{A})
$$

Proof. Since $\mathbf{B}$ is full row rank, then: $\mathcal{R}(\mathbf{B}) = \{\mathbf{y} : \mathbf{y} = \mathbf{B}\mathbf{x} \mid \mathbf{x} \in \mathbb{R}^p\} = \mathbb{R}^n$

$$
\begin{aligned}
\Rightarrow \mathcal{R}(\mathbf{A}\mathbf{B}) &= \{\mathbf{y} : \mathbf{y} = \mathbf{A}\mathbf{B}\mathbf{x} \mid \mathbf{x} \in \mathbb{R}^p\} = \{\mathbf{y} : \mathbf{y} = \mathbf{A}\mathbf{z} \mid \mathbf{z} \in \mathcal{R}(\mathbf{B})\} \\
&= \{\mathbf{y} : \mathbf{y} = \mathbf{A}\mathbf{z} \mid \mathbf{z} \in \mathbb{R}^n\} \\
&= \mathcal{R}(\mathbf{A})
\end{aligned}
$$

∎

2.3 Least square problem

Lemma 2.3 (Least square problem)**.**

$$\min_{\mathbf{x}\in\mathbb{R}^n} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|^2$$

Number of ways to solve lease square problem:

- *Pseudo-inverse*
- *Orthogonal projection*
- *Moore-Penrose inverse*

Solutions. Firstly, let $\mathbf{x}_{\mathrm{LS}}$ is solution, then it must satisfy

$$\mathbf{A}^T\mathbf{A}\mathbf{x}_{\mathrm{LS}} = \mathbf{A}^T\mathbf{y}$$

1. Pseudo-inverse. If $\mathbf{A}$ is full column rank, then solution is unique

   $$\mathbf{x}_{\mathrm{LS}} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{y}$$

   Based on that, some defintions are arised:

   - Pseudo-inverse of $\mathbf{A}$ is $\mathbf{A}^\dagger = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T$
   - Project matrix of $\mathbf{A}$ is $\mathbf{P}_\mathbf{A} = \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T = \mathbf{A}\mathbf{A}^\dagger$
   - Projecting $\mathbf{y}$ onto $\mathbf{A}$ is vector $\Pi_\mathbf{A}(\mathbf{y}) = \mathbf{P}_\mathbf{A} \times \mathbf{y}$

   Note that these definitions above valid only if $\mathbf{A}^T\mathbf{A}$ is invertible, which requires $\mathbf{A}$ full column rank as stated in 2.1

2. If $\mathbf{A}$ is full row rank.

3. If $\mathbf{A}$ is rank deficient.

■

2.4 Big 0, Small o

- Big O $f(x) = O(x)$ if

  $$\lim_{x\to 0} \frac{f(x)}{x} < \infty$$

  Or we can say: $f(x)$ approaches 0 as least as fast as x

- Small o

  $f(x) = o(x)$ if

  $$\lim_{x\to 0} \frac{f(x)}{x} = 0$$

  Or we can say: $f(x)$ approaches 0 faster than x

2.5 First-order necessary condition

Lemma 2.4. *If $\mathbf{x}^*$ is local solution of f over $\mathbf{\Omega}$, then for any feasible direction $\mathbf{d}$, we have:*

$$\nabla f(\mathbf{x}^*)\mathbf{d} \geq 0$$

Proof 1: Taylor approximation. Let $\mathbf{x} = \mathbf{x}^* + \alpha\mathbf{d}$ where $\alpha > 0$, then Taylor expansion gives us:

$$f(\mathbf{x}) = f(\mathbf{x}^* + \alpha\mathbf{d}) = f(\mathbf{x}^*) + \alpha\nabla f(\mathbf{x}^*)\mathbf{d} + o(\|\alpha\mathbf{d}\|)$$

Given α small enough, then:

$$f(\mathbf{x}^* + \alpha\mathbf{d}) \approx f(\mathbf{x}^*) + \alpha\nabla f(\mathbf{x}^*)\mathbf{d}$$

The fact that $\mathbf{x}^*$ is local solution leads to:

$$\begin{aligned} f(\mathbf{x}^*) &\leq f(\mathbf{x}) \\ \Leftrightarrow f(\mathbf{x}^*) &\leq f(\mathbf{x}^*) + \alpha\nabla f(\mathbf{x}^*)\mathbf{d} \\ \Leftrightarrow \nabla f(\mathbf{x}^*)\mathbf{d} &\geq 0 \end{aligned}$$

■

Proof 2: Derivative. Let $g(\alpha) = f(\mathbf{x}^* + \alpha\mathbf{d})$, then

- $g(0) = f(\mathbf{x}^*)$ is local solution
- $g'(0) = \lim_{\alpha\to 0}(g(\alpha) - g(0))/\alpha$

$$\Rightarrow g'(0)\alpha = g(\alpha) - g(0) \tag{5}$$
$$\Rightarrow f'(\mathbf{x}^*)d = g(\alpha) - g(0) \geq 0 \tag{6}$$

■

Lemma 2.5 (First-order condition of convex function)**.** *First-*

Lemma 2.6 (Global solution of convex function)**.**

2.6 Positive/Negative half-space

Why does it exist?

2.7 Order of convergence

Let sequence of real numbers $\{x_k\}$ converges to x^*. The order of convergent sequence $\{x_k\}$ is a positive number p, such that:

$$0 \leq \overline{\lim_{k\to\infty}} \frac{|x_{k+1} - x^*|}{|x_k - x^*|^p} < \infty$$

The notion $\overline{\lim}$ is limit of supreme.

Note that the order of convergence only concerns with the tail of the sequence, as we take limit.

Definition 2.1 (Linear convergence)**.** A sequence has the convergence order of unity is call linear convergence. It is too prevailed so that people make it own defintion. If

$$\lim_{k\to\infty} \frac{|x_{k+1} - x^*|}{|x_k - x^*|} = \beta < 1$$

holds, then sequence is said to converge linearly with convergence ratio (rate) β.

If $\beta = 0$, then it is called superlinear, which is faster than linear. Convergence of any order greater than unity is superlinear.

Warning: Convergence order of 1 is not equivalent to linear convergence, because it might be sublinear. Take sequence $x_k = 1/k$ as an example.

Warning: Superlinear convergence might has the convergence order of unity. Take sequence $x_k = (\frac{1}{k})^k$ as an example.

Lemma 2.7 (Frobenis norm bounds)**.** *Let* $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{K\times K}$.

$$\|\boldsymbol{AB}\|_{\mathrm{F}} \geq \sigma_{\min}(\boldsymbol{A})\,\|\boldsymbol{B}\|_{\mathrm{F}} \tag{7}$$
$$\|\boldsymbol{AB}\|_{\mathrm{F}} \geq \sigma_{\min}(\boldsymbol{B})\,\|\boldsymbol{A}\|_{\mathrm{F}} \tag{8}$$
$$\|\boldsymbol{AB}\|_{\mathrm{F}} \leq \sigma_{\max}(\boldsymbol{A})\,\|\boldsymbol{B}\|_{\mathrm{F}} \tag{9}$$
$$\|\boldsymbol{AB}\|_{\mathrm{F}} \leq \sigma_{\max}(\boldsymbol{B})\,\|\boldsymbol{A}\|_{\mathrm{F}} \tag{10}$$

Proof. To prove (7),

$$\begin{aligned}
\|\boldsymbol{AB}\|_{\mathrm{F}}^2 &= \sum_{i=1}^{K} \|\boldsymbol{A}\boldsymbol{b}_i\|^2 \\
&= \sum_{i=1}^{K} \left\|\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top\boldsymbol{b}_i\right\|^2 \quad \text{(SVD decomposition of } \boldsymbol{A} \text{ always exists)} \\
&= \sum_{i=1}^{K} \left(\boldsymbol{b}_i^\top\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{U}^\top\right)\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top\boldsymbol{b}_i \\
&= \sum_{i=1}^{K} \boldsymbol{b}_i^\top\boldsymbol{V}\boldsymbol{\Sigma}^2\boldsymbol{V}^\top\boldsymbol{b}_i \\
&= \sum_{i=1}^{K}\sum_{j=1}^{K} (\boldsymbol{b}_i^\top\boldsymbol{v}_j)^2\sigma_j^2 \\
&\geq \sum_{i=1}^{K} \sigma_{\min}^2 \sum_{j=1}^{K} (\boldsymbol{b}_i^\top\boldsymbol{v}_j)^2 \\
&= \sum_{i=1}^{K} \sigma_{\min}^2 \left\|\boldsymbol{b}_i^\top\boldsymbol{V}\right\|_{\mathrm{F}}^2 \quad \text{(surprisingly, this is an important step)} \\
&= \sum_{i=1}^{K} \sigma_{\min}^2 \|\boldsymbol{b}_i\|^2 \\
&= \sigma_{\min}^2 \|\boldsymbol{B}\|_{\mathrm{F}}^2
\end{aligned}$$

To prove (9),

$$
\begin{aligned}
\|\boldsymbol{A}\boldsymbol{B}\|_{\mathrm{F}}^2 &= \sum_{i=1}^{K} \|\boldsymbol{A}\boldsymbol{b}_i\|^2 \\
&= \sum_{i=1}^{K} \left\|\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top \boldsymbol{b}_i\right\|^2 \quad \text{(SVD decomposition of } \boldsymbol{A} \text{ always exists)} \\
&= \sum_{i=1}^{K} \left(\boldsymbol{b}_i^\top \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{U}^\top\right) \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top \boldsymbol{b}_i \\
&= \sum_{i=1}^{K} \boldsymbol{b}_i^\top \boldsymbol{V}\boldsymbol{\Sigma}^2\boldsymbol{V}^\top \boldsymbol{b}_i \\
&= \sum_{i=1}^{K}\sum_{j=1}^{K} (\boldsymbol{b}_i^\top \boldsymbol{v}_j)^2 \sigma_j^2 \\
&\le \sum_{i=1}^{K} \sigma_{\max}^2 \sum_{j=1}^{K} (\boldsymbol{b}_i^\top \boldsymbol{v}_j)^2 \\
&= \sum_{i=1}^{K} \sigma_{\max}^2 \left\|\boldsymbol{b}_i^\top \boldsymbol{V}\right\|_{\mathrm{F}}^2 \quad \text{(surprisingly, this is an important step)} \\
&= \sum_{i=1}^{K} \sigma_{\max}^2 \|\boldsymbol{b}_i\|^2 \\
&= \sigma_{\max}^2 \|\boldsymbol{B}\|_{\mathrm{F}}^2
\end{aligned}
$$

The inequality (8) and (10) hold due to the symmetric role of $\boldsymbol{A}$ and $\boldsymbol{B}$. ■

Lemma 2.8 ([2])**.** *Let* $\boldsymbol{X}, \boldsymbol{\Delta} \in \mathbb{R}^{m\times n}$,

$$|\sigma_i(\boldsymbol{X}+\boldsymbol{\Delta}) - \sigma_i(\boldsymbol{X})| \le \|\boldsymbol{\Delta}\|_2 \quad (\le \|\boldsymbol{\Delta}\|_{\mathrm{F}}), \quad 1 \le i \le \min(m,n).$$

3 Probability and Statistic

4 Statistical Learning

4.1 Rademacher Complexity

Definition 4.1. Rademacher complexity of a set $A \subset \mathbb{R}^n$ is defined as

$$\mathcal{R}(A) \triangleq \frac{1}{n} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sup_{\boldsymbol{a}\in A} \langle \boldsymbol{\sigma}, \boldsymbol{a}\rangle\right],$$

where $\boldsymbol{\sigma} = [\sigma_1, \ldots, \sigma_n]$ are n i.i.d. Rademacher random variables, i.e.,

$$\begin{cases} \sigma_i = -1 \text{ with probability } 0.5, \text{ and} \\ \sigma_i = 1 \text{ with probability } 0.5 \end{cases}$$

This is a very general definition. But to put it into the picture: in most cases, A is the set constructed by applying a set of functions (function class) to a fixed dataset with size n. Note that although our interest is purely the "size" or the complexity of our function class, notion complexity introduced by Rademacher complexity is in relative to a fixed dataset.

In most of the time, the best we can do is upper bound Rademacher complexity. We have 2 main strategies do to that.

- Compute it directly from the definition
- Pealing layer by layer.

For the second strategy, we will need the following lemma.

Lemma 4.1 (Contraction lemma [1] (Lemma 26.9))**.** *For each $i \in [n]$, let $\phi_i : \mathbb{R} \to \mathbb{R}$ be a ρ-Lipschitz continuous function, namely for all $\alpha, \beta \in \mathbb{R}$,*

$$|\phi_i(\alpha) - \phi_i(\beta)| \leq \rho \left|\alpha - \beta\right|.$$

For $\boldsymbol{a} \in \mathbb{R}^n$, define

$$\boldsymbol{\phi}(\boldsymbol{a}) \triangleq [\phi_1(a_i), \ldots, \phi_n(a_n)].$$

Let A be a subset of $\mathbb{R}^n$. Let $\boldsymbol{\phi} \circ A \triangleq \{\boldsymbol{\phi}(\boldsymbol{a}) : \boldsymbol{a} \in A\}$. Then

$$\mathcal{R}(\boldsymbol{\phi} \circ A) \leq \rho \mathcal{R}(A)$$

5 Proof

References

[1] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. 2014.

[2] Hermann Weyl. "Das asymptotische Verteilungsgesetz der Eigenwerte linearer partieller Differentialgleichungen (mit einer Anwendung auf die Theorie der Hohlraumstrahlung)". In: *Mathematische Annalen* 71.4 (1912), pp. 441–479.